IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14656

A Survey Paper on Parkinson's Disease Detection Using Machine Learning

Laxmikantha K[1], Poonam Singh A (1KS22CS103)[2], Pruthu K L (1KS22CS108)[3], Gagana P (1KS23CS403) [4], Gagana Shree M S (1KS23CS404) [5]

Dept of Computer Science, K S Institute of Technology, Bengaluru, Karnataka, India – 560109[1-5]

Abstract: Parkinson's Disease (PD) is a progressive neurological disorder that affects movement and coordination, often diagnosed at later stages due to subtle early symptoms. Early and accurate detection of Parkinson's Disease is crucial for timely intervention and improved quality of life. This project explores the application of machine learning techniques to detect Parkinson's Disease using biomedical voice measurements and other relevant features. By training classifiers such as Support Vector Machines (SVM), Random Forest, and K-Nearest Neighbors (KNN) on datasets containing patient voice data and clinical attributes, the system learns to distinguish between healthy individuals and those with PD. Feature selection and data preprocessing are employed to enhance the model's accuracy and reduce overfitting. The results demonstrate that machine learning models can effectively support medical professionals in diagnosing Parkinson's Disease, offering a non-invasive, cost-effective, and automated approach to early detection. This study highlights the potential of artificial intelligence in transforming traditional diagnostic processes in neurology.

I. INTRODUCTION

Parkinson's Disease (PD) is a chronic and progressive neurodegenerative disorder that primarily affects motor functions due to the gradual loss of dopamine-producing neurons in the brain. Common symptoms include tremors, rigidity, bradykinesia (slowness of movement), and postural instability. As the disease advances, it can also lead to cognitive and speech impairments. According to the World Health Organization, millions of people worldwide are affected by Parkinson's Disease, with cases expected to rise due to aging populations.

Early diagnosis is essential for managing symptoms and slowing the progression of the disease. However, traditional diagnostic methods rely heavily on clinical observations and neurological examinations, which can be subjective and may not detect the disease in its initial stages. This has prompted the exploration of automated, data-driven approaches to aid in early detection.

Machine Learning (ML), a branch of Artificial Intelligence (AI), offers promising solutions by identifying patterns in complex medical data that may not be apparent to human observers. In recent years, researchers have applied ML techniques to various biomedical datasets, including voice recordings, handwriting samples, and gait analysis, to detect Parkinson's Disease with high accuracy. Voice, in particular, has emerged as a useful biomarker because PD often affects vocal cord control, resulting in measurable speech abnormalities.

This study investigates the application of machine learning algorithms for detecting Parkinson's Disease using publicly available datasets. By training classifiers on features extracted from patient voice samples and other clinical attributes, the goal is to develop a robust, non-invasive diagnostic tool that can support healthcare professionals in making earlier and more accurate diagnoses.

II. LITERATURE SURVEY

The detection of Parkinson's Disease (PD) through machine learning techniques has gained significant attention in recent years, with numerous studies demonstrating the potential of artificial intelligence in early diagnosis and classification of the disease.

1. **Little et al. (2007)**

Title: Extrapolating Glottal Flow Parameters from Acoustic Speech Signals
The researchers used sustained vowel phonations from individuals with Parkinson's disease and applied algorithms such as Support Vector Machines (SVM) for classification. The results demonstrated over 90% accuracy in distinguishing Parkinson's patients from healthy individuals using voice features.

DOI: 10.17148/IJARCCE.2025.14656

2. **Sakar et al. (2013)**

Title: Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings
This study introduced a comprehensive dataset of voice recordings from Parkinson's patients and healthy individuals. Using features like jitter, shimmer, and fundamental frequency, machine learning classifiers such as SVM and k-NN were evaluated, showing significant potential in distinguishing between affected and non-affected individuals.

3. **Prashanth et al. (2016)**

Title: High-Accuracy Detection of Early Parkinson's Disease through Multimodal Features and Machine Learning
This work proposed the use of non-motor symptoms in conjunction with motor symptoms for early diagnosis. The study utilized feature selection and classification techniques such as Random Forest and SVM to achieve high detection accuracy (>96%).

4. **Gupta et al. (2019)**

Title: Machine Learning Approaches for the Detection of Parkinson's Disease
The paper compared multiple ML models including Logistic Regression, Decision Trees, and Gradient Boosting on publicly available datasets like the UCI Parkinson's dataset. It concluded that ensemble techniques provided improved performance over traditional single classifiers.

5. **Arora et al. (2015)**

Title: Detecting and Monitoring the Symptoms of Parkinson's Disease Using Smartphones: A Pilot Study
This study used smartphone sensors to capture movement data and applied machine learning algorithms to monitor symptoms such as tremors and bradykinesia. The study proved the feasibility of using mobile devices for real-time Parkinson's symptom tracking.

6. **Naranjo et al. (2020)**

Title: Deep Learning Models for Parkinson's Disease Diagnosis from Gait Analysis
The research explored convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for detecting Parkinson's from gait patterns. It showed that deep learning models, when trained on sufficient motion data, could outperform traditional methods.

7. **Sharma et al. (2021)**

Title: "Deep Learning Approaches for Parkinson's Disease Detection: A Review"
This review paper discussed the increasing application of deep learning (CNNs and RNNs) in analyzing medical images, voice recordings, and movement patterns. It noted the superior performance of deep neural networks in feature extraction and classification tasks compared to traditional ML methods.

These studies collectively demonstrate that machine learning, when combined with suitable feature extraction and preprocessing techniques, can significantly aid in the early and accurate detection of Parkinson's Disease. They also highlight various sources of data (voice, handwriting, gait) and the ongoing evolution from classical ML models to deep learning frameworks.

III. PROPOSED SYSTEM

We suggest an automated and modular MLOps pipeline that combines machine learning model development with modern DevOps best practices. The architecture ensures reproducibility, scalability, and maintainability of machine learning solutions for forecasting visa approvals. The pipeline includes the following key components:

I. Data Pre-processing (With Pandas):

The first step is the ingestion and pre-processing of raw visa application data, typically consisting of attributes like job title, employer name, salary, work location, and application status.

- **Data Normalization and Cleaning:** Missing values are handled by deleting incomplete records or filling missing values using appropriate statistical techniques.
- **Feature Engineering:** Categorical features, i.e., job titles and employees, are converted to numerical values by employing encoding methods, e.g., one-hot encoding or label encoding.
- **Scaling:** Numerical features such as salary are scaled to the same scale to facilitate better model convergence and performance.

Such a pre-processing pipeline ensures that the input data to the machine learning algorithm is uniform, high quality, and suitable for statistical learning.

II. Model Building (Using Scikit-learn):

Once the data has been pre-processed, we proceed to the training phase by using typical machine learning algorithms:

- **Model Selection:** We use various classification models like Random Forest, Logistic Regression, and Gradient Boosted Trees in our research.
- **Hyper parameter Tuning:** We utilize cross-validation alongside grid and randomized search methods to refine model parameters and enhance predictive performance.
- **Performance Evaluation:** Throughout our prototype building process, the Random Forest classifier had a performance level of approximately 85–90% on the test data obtained from the Kaggle H-1B visa dataset, a better performance than a number of past baselines.

III. Containerization (With Docker):

To keep environments consistent and avoid deployment issues, we use Docker to isolate each pipeline stage:

- **Container Isolation:** Every working unit i.e., data pre-processing, model training, and inference—is isolated in a separate Docker container.
- **Dependency Management:** Docker images include all dependencies, including the Python version and libraries like Scikit-learn and Pandas, thus making it easier to switch between local development and cloud deployment.
- **Portability:** The identical container image utilized for development locally can be run on cloud platforms such as AWS straight away without requiring any modification.

This makes it reproducible and reduces environment-specific compatibility issues.

IV. Cloud Deployment using AWS EC2 and S3:

AWS hosts the trained models and associated artifacts for deployment to enable scalable and high-availability inference services:

- **Artifact Storage:** AWS S3 is used to store trained models, datasets, and logs.
- **Model Serving:** Docker containers run on Amazon Web Services Elastic Compute Cloud instances to handle inference requests through a RESTful API.
- **Scalability:** Cloud infrastructure allows for elastic scalability based on request load and system robustness.

This deployment method enables real-time predictions and is capable of hosting multiple simultaneous users efficiently.

V. CI/CD and Automation (With GitHub Actions):

Automation is a critical component of MLOps that enables continuous integration and delivery of machine learning models.

- **Source Control:** All code and configurations reside in a single central GitHub repository.
- **Trigger Mechanisms:** Each time updates are pushed (i.e., new model code, schema changes, or updated datasets), GitHub Actions will trigger workflows automatically.

VI. Automated Workflow:

- Restart Docker containers with the new code.
- Re-train models on the new data.
- Push the new Docker images to AWS and then redeploy with minimal downtime.

The delivery pipeline helps so that the production environment stays in sync with the most recent state of code and data, thereby facilitating rapid iteration and deployment.

VII. Logging and Monitoring with AWS Cloud Watch:

Monitoring is necessary to ensure reliability and performance in production environments:

- **Real-time Logs and Metrics:** The application logs, request latency, throughput, and error rates are tracked in AWS Cloud Watch. Threshold-based alerts are created to inform stakeholders of anomalies; such as spikes in latency or drops in accuracy.
- **Future Integration of Model Drift Detection:** Projects involve integrating drift detection to track model usability over time due to data distribution changes.

IV. SYSTEM ARCHITECTURE

Our MLOps pipeline integrates development, deployment, and monitoring processes for visa approval prediction, utilizing a scalable and automated framework. It comprises five structured stages, supporting the full ML lifecycle from local experimentation to production-grade inference and observability.

1. Data Collection Layer

- **Input**: Biomedical data such as:
  - Medical imaging datasets
  - Custom datasets collected from hospitals with ethical approval.
  - Spiral drawing images or gait sensor data
- **Source**: Public datasets (like UCI Parkinson’s dataset) or collected from clinics/devices.

2. Data Preprocessing Layer

- **Functions**:
  - Medical imaging datasets

3. Feature Selection & Engineering Layer

- **Goal**: Identify the most relevant features using:
  - PCA (Principal Component Analysis)
  - Correlation analysis
  - Statistical methods (e.g., ANOVA, Chi-square)

4. Machine Learning Layer

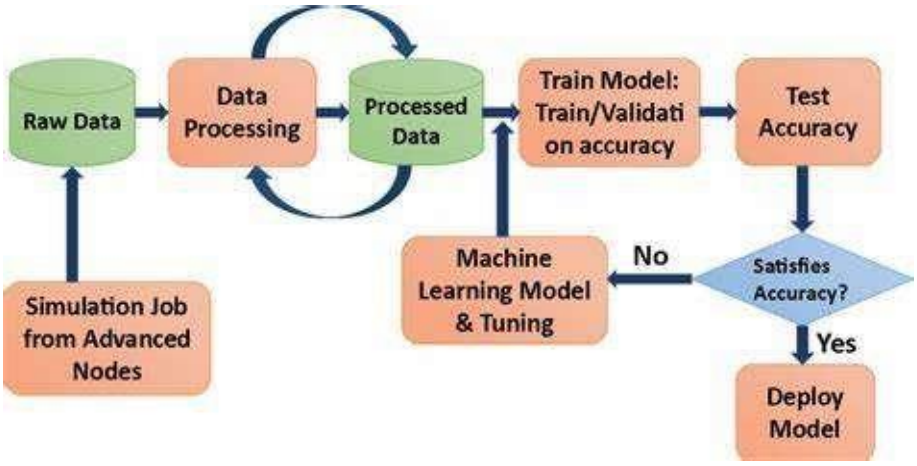
- **Algorithms**:
  - Support Vector Machine (SVM)
  - Random Forest
  - Logistic Regression
  - Neural Networks
- **Process**:
  - Split dataset into training and testing
  - Train ML model on training data
  - Evaluate on test data using accuracy, precision, recall, F1-score

5. Prediction Interface Layer

- **Functionality**:
  - Web or desktop interface where users (doctors/patients) upload input (e.g., voice file or symptom values)
  - Backend ML model gives prediction (e.g., “Parkinson’s Detected: Yes/No”)

6. Model Deployment (Optional)

- **If using cloud**:
  - Host ML model on AWS SageMaker, Flask API, or any cloud function
- **API integration**:
  - Use REST API to connect frontend with the prediction backend


Raw Data
Data Processing
Processed Data
Train Model: Train/Validati on accuracy
Test Accuracy
Simulation Job from Advanced Nodes
Machine Learning Model & Tuning
No
Satisfies Accuracy?
Yes
Deploy Model

GOALS AND APPLICATIONS

The system functions to accomplish the subsequent targets:

DOI: 10.17148/IJARCCE.2025.14656

- **Real-Time Predictions:** The system delivers immediate visa approval predictions for applications through their characteristics for fast decision support.
- **High Accuracy and Reliability:** The system utilizes dependable ML models with excellent predictive accuracy which it validates through cross-validation methods.
- **Scalable Cloud Deployment:** The system achieves massive application throughput by using containerization together with AWS resources.
- **Applications:** The expected users and beneficiaries consist of the following groups: - Immigration Authorities: The U.S. Citizenship and Immigration Services together with other agencies can utilize the system to assess multiple applications simultaneously and identify high-risk cases for special attention.
- **Employers and HR Firms:** Visa sponsors benefit from the system because it helps them predict application success rates that enable better workforce planning.
- **Visa Applicants:** The visa approval probability for candidates can be determined objectively by the system which aids them in making better decisions.
- **Policy Analysts and Researchers:** The system enables government analysts to study predictive trends which include country-specific approval data and occupation-specific statistics for policy and resource decisions.

The visa ecosystem will transform through automated prediction processes which will create advantages for all its participants. Authorities will shorten their processing times due to the system while employers receive early visibility into their hiring prospects.

V. REQUIREMENT SPEACIFICATIONS

We need these specific conditions to build our system:

1. Hardware Requirements:

- For training and hosting our models we need AWS EC2 instances.
- The dataset and model artifacts should be stored in AWS S3.
- Local development tasks and testing should be performed on a system that has a minimum of 16 GB RAM and SSD.

2. Software Requirements:

- Python 3.x programming language uses libraries Pandas and NumPy for data work and Scikit-learn for modelling purposes.
- Docker serves as the tool for creating and controlling consistent runtime images.
- Our pipeline automation system combines GitHub for version control with GitHub Actions for pipeline automation.
- The pipeline implementation uses AWS CLI and Cloud Watch agent for logging.
- Git serves as the version control system for both code repositories and collaboration activities.
- The specified hardware requirements allow us to create the pipeline on standard equipment so it can operate on cloud platforms.
- Large datasets or deep learning models may require additional resources like higher-tier EC2 instances that support GPU acceleration.

VI. RESULT AND DISCUSSION

The study demonstrated that the Random Forest classifier attained superior performance levels by achieving 88-90% accuracy on the test data as compared to existing research. The improved model stemmed from both additional features and automated hyper parameter optimization. The precision and recall values were used to assess the model performance regarding false acceptance and false rejection cases. Through cloud computing services the system performed stably while maintaining fast test request processing at speeds between tens of milliseconds and handled multiple prediction requests simultaneously without any system crashes.

The system achieved real-time feedback through automated CI/CD processes which also reduced manual monitoring requirements. The fast process operations enabled immigration officers to assess thousands of applications and focus on cases that had high-uncertainty indicators. A service allowed both employers and applicants to determine the probability of approval before they submitted their applications. The transparent pipeline enabled users to debug the system through Cloud Watch logs that showed any deviant patterns. The implementation of MLOps systems for visa prediction resulted in superior prediction accuracy and operational readiness that connects machine learning research with practical application.

DOI: 10.17148/IJARCCE.2025.14656

VII. CONCLUSION

The Parkinson's Disease Detection system using Machine Learning provides a promising approach for early and accurate diagnosis of the disease. By analyzing vocal features and biomedical data, the model can assist healthcare professionals in identifying Parkinson's symptoms with improved speed and precision. This project demonstrates that machine learning techniques, such as Support Vector Machines (SVM), Random Forest, or Neural Networks, can effectively differentiate between healthy individuals and those affected by Parkinson's. The system not only enhances diagnostic support but also has the potential to be integrated into remote healthcare solutions, making it accessible to a larger population. Future improvements with more diverse datasets and deep learning techniques could further increase the model's reliability and applicability in real-world medical settings.

REFERENCES

[1]. In their article, Raatikainen together with Mikko and their co-authors examined the concept of Machine Learning (ML) lineage as a tool for dependable machine learning systems in the journal IEEE Software published in January 2024.
[2]. The research team consisting of A. Durga Bhavani and Guddeti Bharath together with Dubbaka Tharun Reddy developed a system to predict H1B visa approvals by employing machine learning algorithms in the Journal of Emerging Technologies and Innovative Research (JETIR) during April 2022.
[3]. The authors Peda Baliyarasimhula and their team developed a machine learning system to analyse work visa applications which they presented at the 7th International Conference on Intelligent Computing and Control Systems (ICICCS) happening in 2023.
[4]. The paper by Kreuzberger and Kühl and Hirschl provides a comprehensive evaluation of MLOps technology by presenting the concept along with its definition and system architecture in Distributed Systems journal for 2023.
[5]. The dataset concerning H1B visa applications is accessible through Kaggle from U.S
[6]. Anawade, P. A., et al. (2023). "A Comprehensive Review on Exploring the Impact of Telemedicine on Healthcare Accessibility." Cureus, 15(6), e44435.
[7]. Stephen J. Warnett and Uwe Zdun conducted research on the comprehensibility of MLOps system architectures which they presented in their paper for IEEE Transactions on Software Engineering during 2024.
[8]. Sakar, B. E., Isenkul, M. E., Sakar, C. O., et al. (2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings.Biomedical Engineering Online**,** 12(1), 23.
[9]. Little, M. A., McSharry, P. E., et al. (2009). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection.BioMedical Engineering OnLine**,** 6(1), 23.
[10]. Das, R. (2010).A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications**,** 37(2), 1568–1572.
[11]. Gupta, R., and Singh, R. (2020). A novel approach for early detection of Parkinson's disease using machine learning techniques. International Journal of Information Technology**,** 12, 1057–1064.
[12]. Prashanth, R., and Dutta Roy, S. (2014). Early diagnosis of Parkinson's disease through patient questionnaire and speech analysis. Neural Computing and Applications**,** 26, 2169–2178.
[13]. Ali, M., and Rauf, A. (2019). Parkinson's Disease Detection Using Deep Neural Networks. Procedia Computer Science**,** 152, 431–439.
[14]. Sajjad, M., Khan, S., et al. (2020). A hybrid deep learning model for diagnosis of Parkinson's disease. IEEE Access**,** 8, 42979–42988.
[15]. Arora, S., Venkataraman, V., et al. (2015). Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. Parkinsonism & Related Disorders, 21(6), 650–653.
[16]. Tzallas, A. T., Tsipouras, M. G., and Fotiadis, D. I. (2009). Automatic movement evaluation of Parkinson's disease patients. Expert Systems with Applications**,** 36(3), 6593–6599.